DOCUMENT RESUME

ED 425 168 TM 028 934

AUTHOR Ruiz-Primo, Maria Araceli; Shavelson, Richard J.; Baxter, Gail P.

TITLE An Approach to Formative Evaluation for Teacher Enhancement Programs.

PUB DATE 1993-04-00

NOTE 65p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS Elementary Secondary Education; Evaluation Methods; Evaluation Utilization; *Formative Evaluation; *Inservice Teacher Education; *Performance Based Assessment; *Professional Development; Program Development; *Program Evaluation

IDENTIFIERS National Science Foundation

ABSTRACT

This paper presents one possible approach to the formative evaluation of a Teacher Enhancement Program (TEP). The approach was applied to a National Science Foundation TEP designed to enhance teachers' knowledge and use of performance assessment technology. The study demonstrates the applicability of the approach to formative evaluation and provides a concrete example of how program development and findings from formative evaluation can be linked together. According to the approach, the study: (1) conceives a TEP as a system of interrelated components that goes through different stages of development; (2) considers the formative evaluation as an iterative process wherein the nature and purpose change with the development of the program; (3) shows how the iterative process renders cumulative knowledge about the program, and how this knowledge may lead to decisions to arrive at a prototype program with characteristics that increase its potential for success; (4) presents a multimethod approach to conduct formative evaluation and show how to triangulate evaluation findings based on different sources of information; and (5) shows how the adaptability of the evaluator's role and his or her knowledge of the subject matter of the program can make a difference in the impact of evaluation findings. An appendix charts content by goal across program development try-outs. (Contains 4 tables, 9 figures, and 32 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 425 168

An Approach to Formative Evaluation for Teacher Enhancement Programs

Maria Araceli Ruiz-Primo and Richard J. Shavelson
University of California, Santa Barbara

Gail P. Baxter
University of Michigan

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Maria A. Ruiz-Primo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

DRAFT

RUNNING HEAD: Formative Evaluation

Paper Presented at the AERA Annual Meeting
Atlanta, Georgia
April, 1993

TM028934

Abstract

This paper presents one possible approach to the formative evaluation of a Teacher Enhancement Program (TEP). The approach was applied to an NSF TEP designed to enhance teachers knowledge and use of performance assessment technology. This study demonstrates the applicability of the approach to formative evaluation and provides a concrete example of how program development and findings from formative evaluation can be linked together. According to the approach, the study: (a) conceives a TEP as a system of interrelated components which goes through different stages of development; b) considers the formative evaluation as an iterative process wherein the nature and purpose change with the development of the program; (c) shows how the iterative process renders cumulative knowledge about the program, and how this knowledge may lead to decisions to arrive at a prototype program with characteristics that increase its potential for success; (d) presents a multimethod approach to conduct formative evaluation and show how to triangulate evaluation findings based on different sources of information, and (e) shows how the adaptability of the evaluator's role and his or her knowledge of the subject matter of the program can make a difference in the impact of the evaluation findings.

An Approach to Formative Evaluation for Teacher Enhancement Programs

For the last decade science education reform has been a major arena for policy action. The reform addresses fundamental questions such as: What teaching methods enable students to understand the nature and culture of science? (e.g., Hurd, 1986) How can educators foster scientific literacy in students? How can science be related to everyday decision making? and How can science understanding be assessed? The current reform addresses these questions in the following ways: Science instruction should parallel the methods used by scientists to understand the natural world (e.g., Raizen, Baron, Champagne, Haertel, Mullis, & Oakes, 1989). From this perspective, students have to do science--observe, hypothesize, record data, make inferences and generalizations--to solve scientific problems. By "doing" science students construct meaning both individually and in groups.

Unless assessment is changed, however, reform in science education will not be comprehensively implemented at the classroom level (e.g., Kulm & Stuessy, 1991; Shavelson, Carey, & Webb, 1990). Consequently, many states have responded with new testing policies which move achievement testing away from multiple-choice tests of basic skills toward performance-based assessments of knowledge and problem solving. Performance assessments, by their nature, are congruent with science curricular reform and constructivist notions of learning on which part of the reform agenda is based (e.g., Shavelson, Baxter, Pine, Yuré, Goldman, & Smith, 1991; Shavelson,

Carey, & Webb, 1990). Typically, students are posed a problem, provided with laboratory equipment, and asked to use these resources to generate a solution. Scores from these performance assessments reflect not only the adequacy of students' solutions but also the procedures used to arrive at their solutions (e.g., Baxter, Shavelson, Goldman, & Pine, 1992).

Changes in the nature and purpose of science instruction and subsequent changes in the nature of assessment exert pressure in the classroom teacher to change instructional and assessment practices (e.g., Shavelson & Baxter, 1990). Teachers are expected to shift from textbook and rote memory to constructivist teaching--teaching based on the students' active construction of knowledge in problem-solving situations. To do this teachers need to be well grounded in science to support an inquiry approach. Teachers need to change their role in the classroom from conveyors of facts and concepts to facilitators of knowledge construction. Teachers need to have skills in managing the physical and social organization of the classroom to support inquiry teaching (e.g., small groups of students working together). Finally, teachers need to have knowledge about the new assessment policies and practices.

To support teachers in the transition from traditional textbook teaching to constructivist teaching a sustained program of in-service education is needed. Such a program would give teachers an opportunity to deliberate about the new perspectives in curriculum, teaching, learning, and assessment (e.g., Hurd, 1986; Shavelson, Copeland, Baxter, Decker, & Ruiz-Primo, in press). In response to this need, the National Science Foundation (NSF) initiated in 1984 a

Teacher Enhancement Program (TEP) to provide effective in-service education and foster the development and dissemination of improved models for conducting in-service education programs for science and mathematics teachers across the country. Seven years later, NSF conducted an evaluation of the TEP to assess its accomplishments so as to determine appropriate future policies, program operation procedures, and funding levels (Fitzsimmons, Carlson, Burnham, Heinig, & Stoner, 1991). Based on a survey of principal investigators who received grants from NSF, the evaluators urged that each TEP funded program "... should be considered as a candidate for formative evaluation early in its life ... and summative evaluation five to seven years after program start-up should be undertaken" (Fitzsimmons et al., 1991, p. xv). The evaluators also recommended that NSF shift its evaluation requirement *from* a summative, self assessment evaluation of program effects *to* a formative evaluation in which data are gathered (Knapp, Shields, St. John, Zucker, & Stearns, 1988) during planning and continually throughout all the phases of program implementation. In short, evaluation information should be used to improve the program as well as to make judgments about it (e.g., Guskey & Sparks, 1991). To date, formative evaluation has been noticeably absent in most of the teacher enhancement programs reported in the literature despite its importance to the development and implementation of an effective TEP (e.g., Knapp et al., 1988).

The purpose of this paper is to present one possible approach to the formative evaluation of TEPs. More specifically, the intent is to: (a) demonstrate how formative evaluation of teacher

enhancement programs might be carried out and how the information it renders can be used for program improvement; (b) provide a concrete example of how to link program development and evaluation; (c) pinpoint a variety of information sources that are useful, understandable, relevant, and practical to program developers; and (d) suggest how evaluators can work together with program developers to improve enhancement programs.

Formative Evaluation of TEPs

Formative evaluation seeks to provide information to guide program improvement and to determine "what works" and "what does not work" during the development and delivery of an enhancement program. Unfortunately, formative evaluation is rarely used during the development of TEPs. Further, if any kind of evaluation is carried out, the information provided is likely to be of tangential interest and utility (e.g., Shavelson et al., 1992). Based on documentation and interviews with NSF principal investigators and participant teachers from exemplary TEPs in science and mathematics, Weiss, Boyd, and Hessling (1990) concluded that the major source of information on the impact of a TEP was anecdotal data (i.e., teachers' reports of their feelings, testimonial letters), and that few projects used systematic classroom observation or provided pretest-posttest data demonstrating gains as a result of participation in the program. Fitzsimmons et al. (1991) arrived at similar conclusions.

Although testimonial information is valuable, it is of limited help in understanding the program and the possible reasons for its effects and consequences. Testimonials such as "Thanks for all you

do" or "The institute helped me to realize that science was all around" (Weiss, Boyd, & Hessling, 1990, pp 2-3), "Teachers, at all levels, tell us of excellent results when they present these demonstrations in their classroom" (McGervey & Heckathorn, 1990, p. 231), and "... anecdotal responses indicate these workshops have been successful" (Sukow, 1990, p. 46) do not tell developers or program administrators what makes the program work or how it can be improved.

Another approach to formative evaluation, common to many NSF projects, is the Concerned-Based Adoption Model (CBAM) (Hall & Loucks, 1978). However, CBAM does not meet the program development and improvement needs of the in-service program itself (e.g., Shavelson et al., 1992). Rather, it focuses on the "adoption" of innovative programs (e.g., Hord, Rutherford, Huling-Austin, & Hall, 1987; Marsh & Sevilla, 1991).

Evaluations conducted by Gayford (1987) and Ellis and Kuerbis (1991) are rare examples in which more systematic formative evaluation data were collected and used to improve TEPs. In both cases the programs' goals were related to science (biotechnology related subjects like genetics and enzymology, and educational computing, respectively) and the evaluations involved repeated trials of the materials and/or the program. Changes to the programs followed from the evaluation results.

Both evaluations included questionnaires to participants and facilitators. However, only Ellis and Kuerbis' evaluation used observation during the workshop, interviews with participants and facilitators, and an instrument to assess the knowledge acquired by

participants as a result of the program. The questionnaires used in both evaluations asked participants about the extent to which their individual requirements were met by the program in terms of information, skills, and the relevance of the materials to the school's needs.

Gayford's evaluation and that of Ellis and Kuerbis illustrate some of the characteristics that should be present in the formative evaluation of a TEP: information beyond anecdotal or testimonial data, revisions of the program through trials, and administration of the program in different sites. The approach to formative evaluation we present incorporates these characteristics in the context of stages of program development. Further, we propose a comprehensive view of formative evaluation at each stage.

An Approach to TEP Formative Evaluation

No single model of formative evaluation exists that can be applied to all enhancement programs. Rather, formative evaluation has to be tailored to the specific needs and circumstances of each enhancement program (e.g. Knapp et al., 1988; Shavelson et al., 1992). In this paper, then, we propose an approach to the formative evaluation of a TEP prior to implementation. We demonstrate through an example how this approach can be used to evaluate a TEP and we suggest how this approach can be generalized to other TEPs.

The approach to formative evaluation *integrates* key ideas in evaluation in such a way that provides a comprehensive characterization of the process of formative evaluation of a TEP. The approach is built on three elements: (a) the TEP program, (b) the

formative evaluation process and (c) the evaluator. Table 1 presents the characterization of this approach to the formative evaluation.

Insert Table 1 Here

TEP Program

For formative evaluation purposes it is useful to identify *stages of maturity* (Cronbach et al., 1980) and *components* of a TEP (Scriven, 1991b). We conceive a TEP as a system of interrelated components which goes through different stages of development.

Developmental stages. An adaptation of the scheme presented by Cronbach et al. (1980) is considered for our purposes. The scheme contains three stages of maturity: the *planned program*--the turning of an idea into a program for action; the *experimental program*--the trial of what the program can accomplish; and the *prototype program*--a preview of what will happen if the practice of the program is made fully-operational (see Table 1, column 1). Although Cronbach et al. present another stage, the established program, when the program becomes established with a permanent budget and an organizational niche, this scheme for formative evaluation seems inappropriate because it assumes that at this point the program has proven effective.

Components. We identify five major interrelated components in TEPs: *context*--nature of the situation in which the program is implemented such as the participants' characteristics, needs, interests; *goals*--the purpose of the program; *materials*--the actual program; *delivery*--carrying out of the program; and *outcomes*--the program's effects (see Table 1, column 2).

The context should guide the selection of goals and these goals should delimit the characteristics of the program, its delivery, and the intended outcomes. Each of these components have subelements. For example, program materials can be characterized as to their content, activities, and sequence which in turn are interrelated.

Program components are frequently of variable quality and the improvement of one or more components leads to the improvement of the program as a whole (Scriven, 1991a; 1992). Some evaluations may focus on one component of the program while others are more systemic (e.g., Shavelson et al., 1992). Viewing the program in terms of its components helps to localize problems and to make specific suggestions (Scriven, 1991a).

Formative Evaluation Process

We view formative evaluation as an iterative process wherein the nature and purpose change with the development of the program. This means that formative evaluation goes through stages too. At each stage different evaluation methods are more or less important depending on the focus. Nevertheless, multiple sources of information, despite the costs, are for the most part preferred, regardless of the stage.

Iterative process. When TEP development is in its early stages, nothing is definite. Evaluative information is used to shape and reshape the program and its delivery. Revisions and modifications are carried out over and over again during a series of reviews and tryouts until a suitable program is developed. Consistent with this view, we postulate that formative evaluation can be characterized as

an iterative process in which the outcome--the program to be implemented--is attained by means of a series of approximations.

This iterative process begins when the very first version of program material is reviewed by a committee of subject matter experts. As a result of this review, the material is revised and a new version is elaborated. Then, the material is piloted in a tryout of the program and new reviews and revisions are carried out. The cycle, review-revision-tryout, is performed over and over again by the program staff until the desired outcome is attained. In this iterative process, the final result is obtained through a series of approximations in which the version of the program obtained on one occasion is always better than the previous versions. Ideally, the iterative process ends when either of two conditions are met: (a) no further modifications are needed, or (b) the benefit of further tryouts do not justify costs.

Reviews and tryouts vary across stages of maturity (see Table 1, column 3). Review of the program by experts may be more necessary at an early than at a late stage, and a tryout with typical consumers in different sites needs fewer modifications if it takes place at an advanced stage of maturity than if it occurs during the planned program stage.

The iterative process renders a cumulative knowledge of the program which increases the program's robustness--its ability to produce the same results with different facilitators, different participants, in different sites, and at different times (e.g., Berk & Rossi, 1990), and the generalizability of the program's effects (e.g., Weiss, 1972). The advantage of accumulating knowledge about a

program is that it provides specific information about conditions under which the program is differentially effective.

Formative evaluation stages. Consistent with the stages of formative evaluation proposed by Scriven (1991b) we postulate the existence of three stages: (1) *In-house reviews*--program evaluation is done primarily by reviewers who are not involved in the development; no tryouts are carried out. (2) *In-house tryouts*--program is evaluated through tryouts with in-house staff and typical participants; reviews by subject-matter experts may be part of the evaluation too. And (3) *Outside tryouts*--the program is tried out and evaluated in different sites with different facilitators and participants (see Table 1, column 4). The third stage of formative evaluation can be considered as an "early-warning summative" evaluation, if a full scale commissioned outside evaluation is carried out.

During the first stage (the planned program), the reviews end when no more modifications to the program can be done on "paper" and tryouts are needed to pilot how effective the program's components are. In-house tryout is the first approximation to the realistic circumstances of the implementation of the program. The delivery is made by the program developers and other staff with typical participants. The evaluation looks for program adaptations (e.g., sequence of content, activities), and for those critical characteristics that are associated with program success. When the best possible adaptations of the program have been done to meet the goals and the conditions of success have been identified, a prototype-program can be tried out in different sites, with different facilitators,

and different participants. This strategy may reveal the range of possible problems in operating the program on a large-scale.

Diversity of methods. A wide range of evaluation methods can be used, from direct observation and/or interviews to quasi- and/or randomized experiments (Shavelson et al., in press). Each method has its strengths and weaknesses. Therefore selection of methods is determined by the type of information needed, who wants/needs the information, and the stage of the program (i.e., planned, experimental, or prototype; see Table 1, column 5). In the planned-program stage, informal evidence is sufficient for developers. Comments by experts on program content and possible difficulties in delivery is what developers need at this stage. In the experimental and prototype stages, more technical evaluations are needed. Evaluations of actual effects of the program are needed and both quantitative and qualitative methods are useful. Small studies can take many forms including quasi-experiments, randomized experiments, and case studies. A combination of methods and sources is needed to cross-check findings and provide a comprehensive perspective of the program. Although redundancy of information has costs, it is as good as getting a second opinion in medical practice (Scriven, 1991a), especially when we use an indicator that is not nearly strong enough to stand by itself. However, qualitative methods become increasingly important when developers want to look at program variations across different sites.

Evaluator

Evaluators themselves are a key factor in the extent to which evaluation information is used (cf. Cronbach et al., 1980; Cronbach,

1982; Weiss, 1988; Patton, 1988; Alkin, 1990). Contrary to popular opinion, and in addition to credibility, two factors may serve to increase the use of formative evaluation information: the adaptability of the evaluators' role and the evaluators' expertise in the subject matter (see Table 1, column 6 and 7).

Adaptable role of a formative evaluator. The traditional evaluator's role is to clearly present the evaluation results so as to increase the probability that the information will be used. In the development of effective TEPs the impact of evaluation findings is crucial (i.e., it is expected that the impact of findings is directly related to program improvement) and the evaluator has to find ways to communicate what the evaluation says about the program. We assume, then, that "impact-of-findings" has to be a major concern for formative evaluators. If the traditional role does not have the desired impact, evaluators have to adopt different roles that lead developers and program administrators to attend to the results.

Ways of communicating may vary according to the characteristics of program administrators and developers, and stages of program maturity. The presentation of evaluation findings may be purely conversational in the planned-program stage. In the experimental- and prototype-program stages, brief reports are presented with technical information. We recommend fast release of evaluation findings, informal conversations about the findings, and brief evaluation reports. Communication overload and delay in release are common faults in evaluation (e.g., Cronbach et al., 1980).

When conversations and open discussions of findings do not have the desired impact, evaluators may look for new ways to be

heard. A possible way to do this is to present suggestions or recommendations together with evaluation findings. This view is contrary to the widely accepted position that the evaluators' role is not to give recommendations, and Scriven's (1992) thesis is an example of this position, "an evaluation without recommendation is like a fish without a bicycle" (p. 43). Nevertheless our experience is that suggestions and recommendations are important in formative evaluations when the goal is program improvement, especially when suggestions and recommendations naturally emerge from the evaluation process. When this is the case, suggestions and recommendations have to be a part of the conversations with developers. In many cases it will take special expertise to set forth worthwhile recommendations (see below).

If suggestions and recommendations do not have the desired impact, evaluators have to modify their role again and look for other ways to be heard. Concrete products may help to call attention to the evaluation results. For example, if the suggestion to change content sequence is not attended to, perhaps the evaluator needs to present a concrete example of a new sequence and justify the changes.

In sum, to say that evaluators have to be creative or have a variety of alternatives at their fingertips to call attention to the evaluation results is to say that they have to learn how to relate to developers and program administrators. "Nothing makes a larger difference in the use of evaluations than the personal factor" (Cronbach et al., 1980. p 6).

Need for subject matter expertise. Evaluators' expertise in the subject matter of the program is helpful in at least two ways: (a)

they can judge the content of the program with a more critical eye than an evaluator who does not have deep knowledge of the subject matter, and (b) their knowledge leads to suggestions and recommendations in a natural way.

It is not uncommon for program administrators to hire experts in the development of instructional materials for a TEP who may or may not know the specific content (e.g., genetics). In either case, an evaluator with subject matter knowledge offers an informed perspective on the program which may facilitate discussions with program administrators and developers. If developers know the subject matter, evaluators may provide another perspective on the program and help in select instructional alternatives. If developers do not know subject matter, evaluators may help developers to conceptualize the program by communicating what might otherwise be overlooked or wrongly perceived (Cronbach, 1982).

In this part of the paper we presented the characteristics of our approach to formative evaluation. The intent was not to provide specific methods for carrying out formative evaluation, but a general strategy. In what follows we apply this approach to an NSF sponsored TEP designed to transfer new assessment technologies to teachers.

Application of the Formative Evaluation Approach

The demand for new assessment technologies that are aligned with science educational reform has lead to an array of alternatives for assessment in science. Hands-on science curricula and science performance assessments are becoming increasingly available. However, many teachers still evaluate students with multiple-choice

tests, creating a mismatch between instruction and assessment at the classroom level (e.g., Shavelson & Baxter, 1990). How can science teachers be helped to learn about and to use alternative assessments? If they change their assessment practices, how will their teaching be affected? Enhancement programs that provide teachers with knowledge and skills necessary to implement performance assessment technologies embedded within the curriculum are needed.

As a part of a project funded by NSF (Shavelson & Baxter, 1990), a team at the University of California, Santa Barbara and the University of Michigan is developing two TEPs to transfer performance assessment technology to teachers and other educators. This paper describes the formative evaluation of one of those TEPs.

The TEP is part of a larger project (Shavelson & Baxter, 1990) devoted to: (a) capturing the new technology involved in developing science performance assessments; (b) providing teachers and other educators with the knowledge and skills needed to understand, select, and use performance assessments embedded within the curriculum, and (c) training teachers and other educators to create and evaluate performance assessments.

The project is organized in an overlapping sequence of three Phases: Performance Assessment Technology, Training Development, and Field Test. In phase I, Performance Assessment Technology, the emerging technology of creating performance assessments, is studied. The goals are to produce and evaluate performance assessments to be used as part of teacher pre-service and in-service education, and

to make explicit the new technology's concepts and procedures so they can be transferred to teachers and other educators.

In phase II, Training Development, a two level system of teacher enhancement is being developed. Level I training provides pre- and in-service teachers with knowledge and skills to understand, select, and use performance assessments. Level II training provides district personnel, teachers, and other educators with the knowledge and skills needed to create and psychometrically evaluate alternative assessments. The major activities in Phase II are the development and evaluation of Level I and Level II training, culminating in prototypes for field testing.

Phase III, Field Test, involves field testing both the Level I and Level II prototypes to evaluate how well the training can be implemented in school districts with hands-on elementary science curricula, and the degree to which training meets its goals.

The project has two Principal Investigators with extensive experience in the development and evaluation of performance assessments. For the development of Level I and Level II training the project has two instructional developer with extensive experience in developing teacher enhancement programs.

During the development of the TEP for Level I training a formative evaluation was carried out following the approach described above. This paper focuses on the second stage of development--in-house tryouts with the experimental program over a one-year-and-a-half period.

The TEP Program

Characteristics. The goals of the TEP were to provide pre- and in-service teachers with the knowledge and skills to: (1) understand the nature of assessment reform, (2) select existing assessments that are appropriate for evaluating individual student achievement or for monitoring the curriculum, and (3) use these assessments in their classrooms (Shavelson & Baxter, 1990).

After reviews, conversations, and revisions, the planned program addressed four basic content topics: (a) knowledge about performance assessments, (b) the nature of a good performance assessment, (c) relationship between performance assessment and the curriculum, and (d) the skills needed for assessing performance. The activities had participants conduct two hands-on experiments and use their corresponding scoring systems, as well as conduct one of the experiments with a computer simulation.

The program can be characterized as a *training approach* to staff development (Sparks & Loucks-Horsley, 1990). First, it is a workshop-type program in which the facilitator is the expert who establishes the content and flow of activities. Second, the training sessions are conducted with a clear set of objectives for learner outcomes. Third, the facilitator's role is to select those activities that will aid teachers in achieving the desired outcomes. This training approach has been considered useful when outcomes like awareness, knowledge, and skills development are considered, or when teachers require demonstrations and practice of instructional skills or techniques to fully understand their implementation (Spark & Loucks-Horsley, 1990).

A major goal during the development of the Level I training program was to achieve a *prototype training program* that could be "exported" to other trainers. In other words, the dissemination of the Level I training program was based on the creation of a package which could be sent to school districts and be used by different trainers in different sites.

The Formative Evaluation Process

Figure 1 summarizes the process of the formative evaluation proposed for the Level I training program. We limit ourselves to the second stage of formative evaluation (in-house tryouts of the experimental program), first, because the audience is familiar with the in-house review process--the discussions, the reviews, and the paper work--that take place during the development of program; second, because the evaluation of the prototype program (Phase III of the project) is still in progress.

According to our approach, during the second stage of the evaluation process in-house tryouts were carried out with developers or project staff as facilitators and typical participants (see Figure 1).

Insert Figure 1 Here

The evaluation focused on three components of the experimental-program: program materials, delivery, and outcomes (see Table 1, column 2). Context and goals were not evaluated for two reasons: First, evidence already existed to indicate that TEPs on the use of alternative assessments in classrooms was strongly needed (e.g., Kulm and Stuessy, 1991; Shavelson & Baxter, 1990; Shavelson,

et al., 1992). Second, it was assumed that the evaluation of the adequacy of the goals had been made by the NSF in considering the funding of this TEP.

To describe the process of evaluation in the jargon used in the field of staff development, hereon 'workshop' will refer to the TEP evaluated.

Formative evaluation questions. Evaluation focused on three major characteristics of the workshop program: (1) *Workshop Materials*: "Are the workshop's content and activities coherent and likely to attain the program's goals?"; (2) *Workshop Delivery*: "Which elements in the workshop's delivery lead to the accomplishment of goals?"; and (3) *Workshop Outcomes*: "Do the participants' knowledge and skills actually change as a result of the workshop?"

The workshop materials were evaluated in terms of the program's goals: understanding, selecting, and use of performance assessments. The material was reviewed and revised in light of these goals and the information (content and hands-on experiences) that Principal Investigators (PIs) thought should be included in the program to achieve them. Figure 2 shows the content for each goal. The working philosophy during the development of the program was that hands-on experiences were the best way to communicate to teachers and other educators the workshop content.

Insert Figure 2 Here

Evaluation of workshop delivery focused on characteristics of the workshop "facilitators" and instructional methods. For example, information was collected on: (a) knowledge about the content, and

(b) instructional methods such as mini-lectures, group discussions among participants, and the material such as artifacts--notebooks, transparencies, and wall posters, and the hands-on activities.

As outcomes, the purpose of the Level I training workshop was to provide participants with the opportunity to "become familiar" with, not "become experts" in the nature, selection and use of performance assessments.

Formative evaluation design and instruments. Five tryouts and reviews of the workshop program during its development were carried out. According to our approach, the tryouts are viewed as iterative pilot studies. The formative evaluation, then, took place before and after the delivery of the workshop, as well as in a pretest-posttest design during the delivery. With this design, at each stage of each tryout the information provided by the evaluation directly impacted the workshop material and delivery. To provide a comprehensive perspective of the workshop and to cross check findings, different sources and methods of data collection--direct observation, questionnaires, analysis of documents, participants' products, and interviews with facilitators--were used.

Table 2 presents an integrated perspective of the design, the sources of information and the methods of data collection for the formative evaluation. The design takes into account the three opportunities to collect formative evaluation information for each tryout--before, during, and after delivery of the workshop, and the instruments used to collect the data.

Insert Table 2 Here

Prior to each workshop, and based on the evaluation of the preceding tryout, workshop material was reviewed and revised. On the second and the fourth tryout, program developers completed the Developers Questionnaire about their perceptions regarding the opportunity given to participants to learn the content and the participants' expected level of knowledge as a consequence of the workshop. The instrument was applied only when substantial changes to the materials were done. This information was used during the discussions related to the content of the workshop.

During the delivery of the workshop, facilitators and participants were observed by the evaluator. In addition, as part of the workshop, participants had to use scoring forms. These participants' scoring forms were examined and compared with an expert's score form (the criterion) for the same students.

Information on the participants' knowledge of the content was evaluated by the Self-Report Knowledge Inventory (SRKI) in a pretest-posttest design. The SRKI is a self-rating instrument that asks participants to provide information about their knowledge on specific topics of the program. Self-rating instruments of this kind do not have the appearance of an achievement test, may be highly correlated with actual performance, and take only a short time for responding (see Tamir & Amir, 1981; Young & Tamir, 1977).

At the end of the workshop, participants completed two questionnaires. The Opportunity to Learn Questionnaire renders information on the opportunity that participants think they had to learn the topics of the workshop. The Opinion Questionnaire provides information on the content, activities, and delivery of the

workshop. Open-ended questions about benefits and limitations of the content, and about the organization of the program were included in the questionnaire.

In the last two tryouts, Facilitators were interviewed about the content and the delivery after the workshop. An Administration of Performance Assessment Questionnaire was given to participants who used performance assessments in their classrooms. In this questionnaire participants reported on rating scales and in written comments what they did when they used performance assessments.

In-House Tryouts. The program was tried out on four occasions with developers and project staff as facilitators and teachers from an NSF funded project, the Science for Early Educational Development (SEED) as participants. SEED was chosen because (a) it trains teachers to use a hands-on science curriculum; (b) the performance assessments can be embedded within SEED's hands-on science curriculum; (c) there is a long standing collaboration between PI's and that school district; and (d) there is a district commitment to enhancing quality of science education. The other tryout was carried out with student teachers in UCSB's Teacher Education Program. Table 3 describes the characteristics of the participants. Fifty-eight in-service elementary science teachers and five pre-service teachers participated across the five tryouts.

Insert Table 3 Here

The tryouts 1 through 5 were carried out in sessions of respectively, two, one, two-and-a half, three, and three days. The

first three tryouts were delivered by the developers of the workshop program; the other two by project staff.

The only sources of evaluation information used for the first tryout were direct observation during the delivery of the workshop and review of the workshop materials. For the remaining tryouts, all the sources of information listed in Table 2 were used when possible, since the application of some instruments depended on conditions external to the evaluation process. For example, questionnaires on the administration of performance assessments in classrooms were not given to all participants because either performance assessments were not available for every grade, or circumstantial problems (e.g., teachers were sick the day the administration of assessment was scheduled).

Evaluation Findings

The evaluation focused on three major questions about the workshop: (1) "Are the workshop's materials--content and activities--coherent and likely to attain the workshop goals?"; (2) "Which elements in the delivery lead to the accomplishment of the workshop goals?"; and (3) "Do participants' knowledge and skills actually change as a result of the workshop?"

First, we present a summary of the impact of the evaluation findings across time--the major decisions made about the content and delivery of the workshop. The intent is to provide a general picture of results within which to locate the specific examples presented later. Next, we provide two examples of the formative evaluation process according to our approach. The first example addresses program content and the other delivery. Both track the

formative evaluation process across tryouts to demonstrate impact (or lack thereof) on program improvement. These examples also demonstrate how findings were cross-checked with multiple-sources. Finally, we suggest future directions for the program based on evaluation findings to date.

Summary of Evaluation Findings Impact

Several important findings have emerged to date through the iterative tryouts of the workshop. These findings are presented for each of the elements evaluated: Workshop Materials, Workshop Delivery, and Workshop Outcomes. Then, a brief description of the evolution of the relationship between Evaluators and Program developers is presented as well.

Workshop materials. Over the tryouts, workshop content and activities were changed based on evaluation results. The quantity of information presented in the workshop has increased as has the accuracy of the content, especially on reliability, validity, and utility of performance assessments (Appendix 1 presents the workshop content for each tryout).

The variety of hands-on activities increased by giving participants the opportunity to try different types of performance assessments (e.g., comparative, decompositional, and taxonomical tasks), different methods (e.g., from observation of hands-on performance to notebooks, to computer simulations, to paper-and-pencil alternatives). Moreover, the number of exercise increased, permitting teachers to apply the knowledge acquired during the workshop to real-classroom-life situations involving the use and selection of performance assessments.

Supplementary information was written by the evaluators and research articles were offered to facilitators to provide a background on the technical characteristics of performance assessments.

Workshop delivery. Two elements in the delivery were found to be critical in meeting workshop goals. The first was Facilitators' knowledge about performance assessments, and their experience with the hands-on assessments used in the workshop. Evaluation findings made evident that workshop content (e.g., information on the psychometric characteristics of performance assessments, hands-on science teaching, administering and scoring assessments) was too complex for Facilitators to read and learn from a manual and present them to teachers.

An important finding, based on direct observation of workshop delivery across the tryouts, was that the Facilitators' Manual helped Facilitators "to know the content" but not "to own the content" (e.g., to discriminate when a discussion has to be brought to a close, or how to direct the discussion to meet workshop goals, or what to respond to "non-scripted questions"). Furthermore, experience with the administration and scoring of performance assessments was found to be an important characteristic that Facilitators needed to "own" to deliver the workshop properly.

Workshop outcomes. The SRKI used in the pretest-posttest design was given for the first time in the second tryout of the workshop. No significant differences between the pretest and the posttest scores were found. By contrast, significant differences were found in tryouts 3, 4, and 5. Table 4 presents the descriptive

statistics and the reliability coefficients (i.e., internal consistency) of pretest and posttest across the tryouts.

Insert Table 4 Here

Relationship between the Evaluators and Program Developers. The relationship has evolved, to date, in four stages: (1) extensive discussions around the evaluation results with recommendations for specific changes--a strategy that failed to produce desired changes; (2) provision of selected research articles related to the workshop content to familiarize developers with critical concepts and research findings--this approach did not work; (3) increasing the program developers' hands-on experience with assessing students--an approach that did not work; and (4) providing Program Developers with draft materials to be incorporated directly in the workshop content. This last strategy was found to produce the desired results, and led to modifications in the content and delivery of the program.

Examples of the Formative Evaluation Across Tryouts

One of the purposes of this paper is to demonstrate how the information rendered by our formative evaluation approach can be used for program improvement. Here we present two concrete examples of how we have linked program development and formative evaluation.

Workshop Content. In order to address the goal of familiarizing participants with the use of performance assessments, an essential piece of content deals with the administration, score, and interpretation of performance assessments. Figure 3 presents the

evolution of the content of "interpretation of performance assessments scores" across the five tryouts.

Insert Figure 3 Here

Based on the reviews of workshop materials, the evaluation pointed out that the workshop had no component on interpretation of performance assessment scores. This finding did not have the desired impact for the first four workshop tryouts. As a last resort, the evaluators and project staff wrote draft materials to be incorporated directly into the workshop program. These materials included a strategy for score interpretation accompanied by a hands-on exercise that involved interpreting performance assessments scores for a classroom of students.

The evaluation findings from reviews of the workshop content was cross-checked with other sources of information, for example, with the participants' perceptions on the workshop. Participants were asked to write down what they believed were the workshop's goals (i.e., "What do you think the major goals of the workshop were"?) and the benefits they gained from it (i.e., "Please, list the major benefits of this workshop, and star the most important one") at the end of the last day of the delivery. These written statements were classified by goals--understand, use, and select performance assessments. Figure 4 and 5 present the percentage of participants' statements that addressed each workshop goals and the percentage of statements that addressed benefits of the workshop, respectively. Most of the statements about workshop goals and benefits across the last four tryouts referred to understanding of performance

assessments (e.g., "introduction to the performance assessments"; "to heighten our awareness of the science assessment reform"). Few participants perceived use and selection of performance assessments as important goals or benefits of the workshop.

Insert Figure 4 Here

Insert Figure 5 Here

Another questionnaire item asked, "What recommendations would you make about the organization and content of this workshop that you think would help to improve it?" Figure 6 presents the percentage of participants' comments for improvement of workshop content by goal. Most responses recommended including information on how to use performance assessments in their classrooms (e.g., "more focus on how to actually use performance assessments with all the kids in a regular classroom"; "more opportunity for application possibilities in the classroom").

Insert Figure 6 Here

Participants' perceptions of workshop goals, benefits, and content reflected the same findings as the review of the material: the need to include more information on how to use performance assessments in classrooms. Since participants are teachers, clearly it is important to provide them with information on how to administer, score, and interpret performance assessments in their classrooms.

Workshop Delivery. This example deals with the delivery of one of the hands-on activities: The Paper Towels Assessment. The

purpose of inserting this activity in the workshop was twofold for participants: to directly experience a performance assessment, and to help participants realize the importance of using standardized scoring forms with performance assessments. This assessment was developed for fifth- and sixth-grade students. They were asked to find out which of three paper towels absorbs, holds, or soaks up the most water and which one the least (see Baxter et al., 1992).

In this activity, participants were divided in groups of four, one member playing the role of a student, and the other three the role of observers who evaluate the "student's" performance (without a scoring form). Facilitators set-up the equipment, gave participants the task, and tracked how each group was doing (e.g., admonishing observers not to help the "student" design the experiment, or not to talk to each other about scoring criteria).

The evaluation findings presented in Figure 7, were based on direct observation on how Facilitators implemented the activity across the five tryouts. The set-up of the activity and directions to participants in the first tryout were inadequate. Consequently, participants were unable to conduct the activity properly which resulted in stereotypic performances and observers' scores that were unrealistic (all "A's").

Based on the implementation of this activity, we learned that although the content of the program may look simple and comprehensible at first glance, it takes a great deal of practice and expertise to deliver it properly. In the last two tryouts, Facilitators were not the developers of the workshop program, but other members of the project staff. These new Facilitators practiced the

administration and the scoring of Paper Towels before these last tryouts, which improved the delivery of this activity.

Insert Figure 7 Here

Although by the third tryout the delivery had improved, to save time, the program developers decided to modify the Paper Towels assessment from its validated form. Instead of using three paper towels, the assessment was implemented with two. This change was more than cosmetic. Since the scoring form was developed for three paper towels the change created a mismatch between the scoring form and the activity. As it can be seen in Figure 7 this mismatch is still an issue.

Participants' perceptions were used to triangulate on the findings from direct observation. Participants were asked two questions: (1) "In your opinion, how helpful were the following activities to understand, use, and select science performance assessments?", and (2) "Do you think you can use performance assessments in your classroom?" In the second and third tryouts, some participants considered the activity "not helpful at all", which corresponded to the evaluation findings on the implementation of the activity obtained from direct observation (see Figure 8). As the implementation of the activity improved, the helpfulness of the activity (Figure 8) and the possibility of using this kind of assessment in class (Figure 9) was more evident for participants. On the fourth and fifth tryouts, most participants considered the activity "helpful" or "very helpful", and from the third tryout on, participants

believed that they could use the performance assessments in their classrooms.

Insert Figure 8 Here

Insert Figure 9 Here

The development of the workshop program and the impact of the evaluation findings during the process of development is full of examples like the two we have presented above. Changes in the content related to the understanding, use and selection of performance assessments, and the adequacy of the delivery of the content has been the "every-day-story" of this formative evaluation process.

Future Workshop Alternatives Based On Formative Evaluation Findings

Issues related to the form of the prototype workshop program are currently being discussed in the project. The nature and extent of the content of the workshop, the training of the Facilitators, and the most appropriate way to disseminate the program are major concerns.

Nature and extent of the workshop. The length of the workshop is currently an important concern. The difficulty of scheduling three-day workshops with teachers during the school year (besides the costs) and the impossibility of doing it with pre-service teachers (eighteen hours for student teachers is approximately half an academic for a 4 unit course) called for a

variety of alternative workshops. The project administrators, in conjunction with NSF project officers, decided to construct three different prototype versions of the workshop program: a three-day workshop, a one-day workshop, and a three-hour workshop. The last two versions will be created from the original workshop.

Facilitators training and dissemination. Facilitators must be knowledgeable and experienced not only in hands-on science education but also in use of performance assessments to adequately conduct the workshop. This finding has impacted the project in at least two ways. First, systematic training for Facilitators is needed. Such training, in our experience, may take more than 50 hours. Second, the original plan to "export" the prototype workshop program as a package to other school districts is now known to be unrealistic. The expertise in both hands-on science and in performance assessments is a very scarce commodity. One alternative has been considered to solve this problem: to transfer the workshop program to private-sector and government-funded organizations that will take responsibility for training teachers and administrators across school districts.

Conclusions

The purposes of this paper were to present an approach to the formative evaluation of TEPs, to demonstrate how this approach can be carried out, and to show how the information it renders can be used. The formative evaluation study we presented:

(1) Demonstrates the applicability of the approach to formative evaluation.

(2) Provides a concrete example of how program development and findings from formative evaluation can be linked together.

(3) Shows how the iterative evaluative process renders cumulative knowledge about the program, and how this knowledge may lead to decisions to arrive at a prototype program with characteristics that increase its potential for success.

(4) Presents a multimethod approach to conduct formative evaluation and shows how to triangulate evaluation findings based on different sources of information.

(5) Shows how the adaptability of the evaluator's role and his or her knowledge of the subject matter of the program can make a difference in the impact of the evaluation findings.

References

Alkin, M. C. (1990). Debates in evaluation. Newbury Park: Sage Publications.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. Journal of Educational Measurement, 29(1), 1-17.

Berk, R. A. & Rossi, P. H. (1990). Thinking about program evaluation. Newbury Park: Sage Publications.

Cronbach, L. J. (1982). Designing evaluations of educational and social programs. San Francisco: Jossey-Bass Publishers.

Cronbach, L. J. & Associates (1980). Toward reform of program evaluation. Aims, methods, and institutional arrangements. San Francisco: Jossey-Bass Publishers.

Ellis, J. D. & Kuerbis, P. J. (1991). A curriculum for preparing science teachers to use microcomputers. School Science and Mathematics, 91(6), 247-254.

Fitzsimmons, S. J., Carlson, K., Burnham, R., Heinig, S., & Stoner, D. (in preparation). A study of NSF Teacher Enhancement Program principal investigators: 1984-1989 (Draft). Cambridge, MA: Abt Associates Inc.

Gayford, C. (1987). Biotechnology 13-18: In-service training for teachers. Journal of Biological Education, 21(4), 281-287.

Guskey, T. & Sparks, D. (1991). What to consider when evaluating staff development. Educational Leadership, 49(3), 73-76.

Hall, G. E. & Loucks, S. (1978). Teacher concerns as a basis for facilitating and personalizing staff development. Teachers College Record, 80(1), 36-53.

Hord, S. M., Rutherford, W. L., Huling-Austin, L. & Hall, G. E. (1987). Taking charge of change. Alexandria, VA: Association for Supervision and Curriculum Development.

Hurd, P. H. (1986). Perspectives for the reform of science education. Phi Delta Kappan, 67(5), 353-358.

Knapp, M. S., Shields, P. M., John, M., Zucker, A. A., & Stearns, M. S. (1988). An approach to assessing initiatives in science education: Summary Report: Recommendations to the National Science Foundation. SRI Project No. 089. Stanford, CA: SRI International.

Kulm, G. and Stuessy, C. (1991). Assessment in science and mathematics education reform. In G. Kulm and S. Malcom (Eds.). Science assessment in the service of reform (pp. 71-87). American Association for the Advancement of Science.

Marsh, D. D. & Sevilla, J. M. (1991). An analysis of the implementation of Project SEED: An interim report. Technical Report, University of Southern California.

McGervey, J. D. & Heckathorn, D. (1990). Retraining teachers in physics. Physics Teacher, 28(4), 230-231.

Patton, M. Q. (1988). The evaluator's responsibility for utilization. Evaluation Practice, 9(2), 5-24.

Raizen, S. A., Baron, J. B., Champagne, A. B., Haertel, E., Mullis, I. V. S., & Oakes, J. (1989). Assessment in science education: The middle years. Washington, DC: The National Center for Improving Science Education.

Scriven, M. (1991a). Evaluation Thesaurus (4th ed.), Newbury Park, CA: Sage Publications.

Scriven, M. (1991b). Beyond formative and summative evaluation. In M. McLaughlin & D. C. Phillips (Eds.). Evaluation and Education: At quarter century (pp 19-64). Ninetieth Yearbook of National Society for the Study of Education. Chicago, IL: The University of Chicago Press.

Scriven, M. (1992). Hard-won Lessons in program evaluation. Unpublished manuscript, Western Michigan University, Evaluation Center.

Shavelson, R. J. & Baxter, G. P. (1990). Transfer new assessment technologies. to teachers and other educators. National Science Foundation Grant No. TEP 90-55443.

Shavelson, R. J., Carey, N. B., and Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. Phi Delta Kappan, 71(9), 692-697.

Shavelson, R. J., Baxter, G. P., Pine, J., Yuré, J., Goldman, S. R., & Smith, B. (1991). Alternative technologies for large scale science assessment: Instrument of education reform. School Effectiveness and School Improvement, 2(2), 97-114.

Shavelson, R. J., Copeland, W., Baxter, G. P., Decker, D. L, & Ruiz-Primo, M. A.(1992). Inservice education models for enhancing the teaching of science. In S. Fitzsimmons (Ed.) Pre-college teacher enhancement in science and mathematics: Status, issues, and problems (in preparation).

Sparks, D. & Loucks-Horsley, S. (1990). Models of staff development. In W. R. Houston, M. Haberman, J Sikula (Eds.) Handbook of research on teacher education (pp. 235-250). New York: Macmilan Publishing Company.

Sukow, W. W. (1990). Physical science workshop for teachers using interactive science exhibits. School Science and Mathematics, 90(1), 42-47.

Tamir, P. & Amir, R. (1981). Retrospective curriculum evaluation: An approach to the evaluation of long term effects. Curriculum Inquiry, 11(3), 259-278.

Weiss, C. H. (1972). Evaluation research. Englewood Cliffs, NJ: Prentice-Hall.

Weiss, C. H. (1988). Evaluation for decisions. Is anybody there? Does anybody care? Evaluation Practice, 9(1), 5-19.

Weiss, I. R., Boyd, S. E., & Hessling, P. A. (1990). A look at exemplary NSF teacher enhancement projects. Chapel Hill, NC: Horizon Research Inc. (Suite 220, 111 Cloister Ct. Chapel Hill, NC 27514, (919), 373-383.

Young, D. B. & Tamir, P. (1977). Finding out what students know. The Science Teacher, 44, 27-28.

Table 1. Characterization of an approach to formative evaluation of TEPs.

TEP PROGRAM		FORMATIVE EVALUATION PROCESS			EVALUATOR	
Stage of Maturity of the Program	Program's Components	Iterativity Process: Conditions	Stage of Formative Evaluation: Focus	Diversity of Methods of Evaluation	Adaptable role	Expert in the subject matter
Planned Program	Context Goals Materials Conditions of Delivery	Reviews- Revisions of the program before any tryout is carried out.	<u>In-house Reviews</u>: Information about the feasibility of the program.	Informal evidence like comments to the content and possible difficulties.		
Experimental Program	Context Goals Materials Delivery Outcomes	Tryouts of the program with in-house staff and typical consumers. Revisions. Reviews are appropriate too.	<u>In-house Tryouts</u>: Information about operation of the program: Variations and characteristics associated with more or less success.	Quantitative (e.g., small studies using quasi experimental or randomized designs) and qualitative methods (e.g., case studies)		
Prototype Program	Context Goals Materials Delivery Outcomes	Tryouts of the program in different sites with similar conditions to the proposed for operation. Revisions. Reviews are less necessary but are appropriate too.	<u>Outside Tryouts</u>: Information on differences in delivery and effects from site to site, possible problems, and costs of implementation.	Quantitative and qualitative methods are appropriate too. Research designs for estimating effectiveness are highly recommended		

Table 2. Design of the formative evaluation of the workshop to transfer performance assessment technology.

			Tryouts				
			Bef-	Delivery			Af-
Focus of the Evaluation	Source of Information	Method of Data Collection	ore	Pre	Dur ing	Post	ter
Workshop Material	Documents	Review of documents.	X				
	Workshop Developers	Developers Questionnaire	X				
	Workshop Participants	Opportunity to Learn Questionnaire				X	
	Facilitators	Interview on workshop content					X
Workshop Delivery	Workshop Participants	Opinion Questionnaire				X	
	Evaluator	Direct Observation during the delivery.			X		
	Facilitators	Interview on workshop delivery					X
Workshop Outcomes	Workshop Participants	Questionnaire about knowledge and skills acquired during the program		X		X	
		Participants' products			X		
		Administration of Performance Assessment Questionnaire					X

Table 3. Characteristics of the workshop participants across tryouts.

Tryouts	1	2	3	4	5
Number of Participants	18	5	13	10	17
Gender					
Females	17	5	11	9	16
Males	1	-	2	1	1
Ethnicity					
Anglo	12	5	7	4	9
Asian	1	-	2	1	0
Black	4	-	3	5	4
Latino	1	-	1	0	4
Grade Level					
Pre-service	-	5	-	-	-
K	3	-	-	-	-
1	3	-	2	1	2
2	3	-	3	2	2
3	4	-	2	3	8
4	3	-	3	2	1
5	2	-	3	2	4

Table 4. Descriptive statistics for the Pretest and the Posttest of each tryout.

	1		2		3		4		5	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
n=	18		5		13		9		17	
Max=	-		60		55		65		48	
Mean	-	-	35.20	38.60	18.78	33.0*	23.66	54.0*	19.76	37.74*
S.D.	-	-	14.04	10.01	8.29	7.87	7.29	3.84	6.61	9.28
Reliability	-	-	-	-	.85	.92	.89	.76	.90	.91

* Significant difference between Pretest and Posttest
- No data

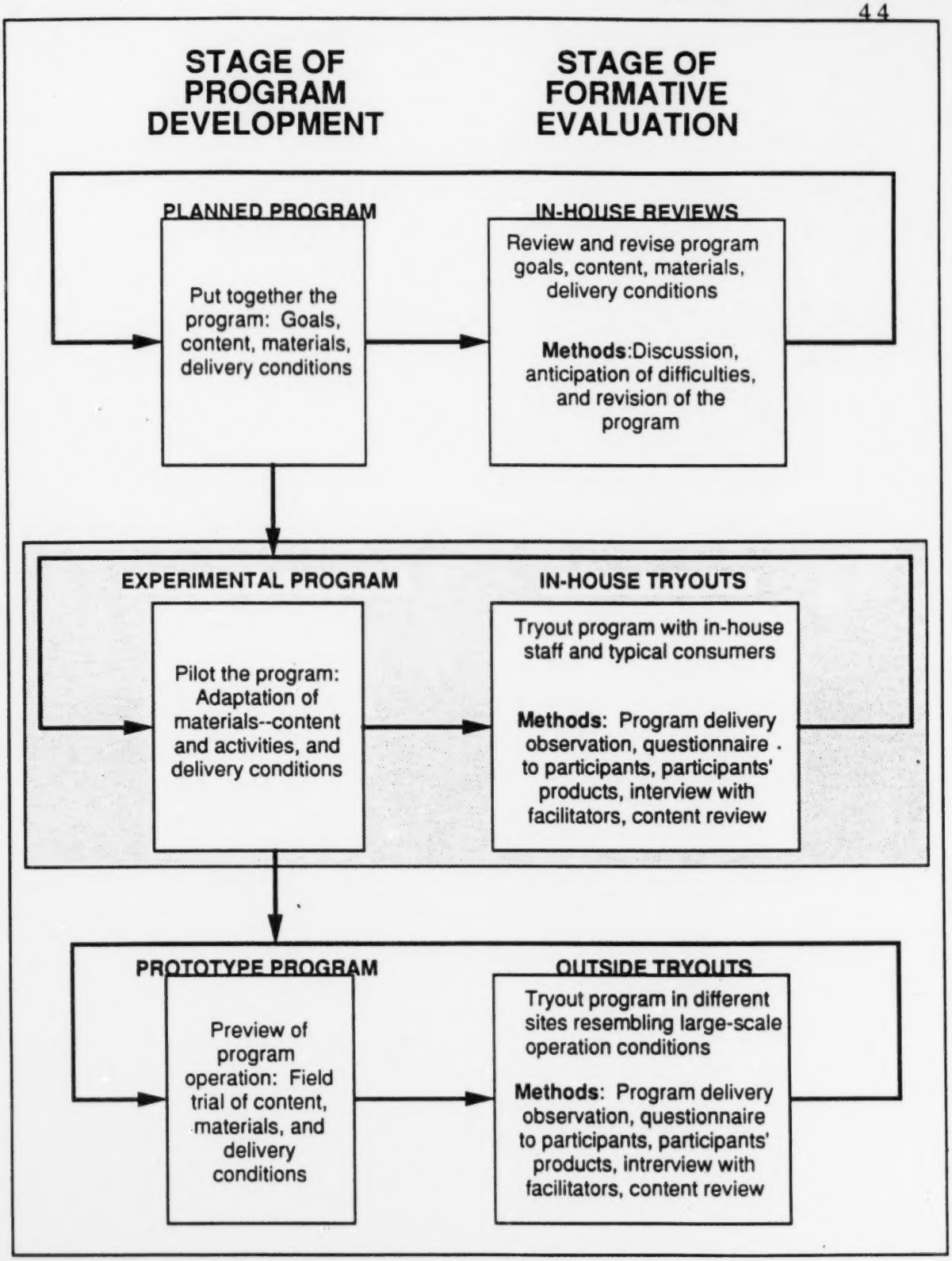
STAGE OF PROGRAM DEVELOPMENT
STAGE OF FORMATIVE EVALUATION
PLANNED PROGRAM
Put together the program: Goals, content, materials, delivery conditions
IN-HOUSE REVIEWS
Review and revise program goals, content, materials, delivery conditions
Methods:Discussion, anticipation of difficulties, and revision of the program
EXPERIMENTAL PROGRAM
Pilot the program: Adaptation of materials--content and activities, and delivery conditions
IN-HOUSE TRYOUTS
Tryout program with in-house staff and typical consumers
Methods: Program delivery observation, questionnaire to participants, participants' products, interview with facilitators, content review
PROTOTYPE PROGRAM
Preview of program operation: Field trial of content, materials, and delivery conditions
OUTSIDE TRYOUTS
Tryout program in different sites resembling large-scale operation conditions
Methods: Program delivery observation, questionnaire to participants, participants' products, intrerview with facilitators, content review

Figure 1. Strategy of formative evaluation

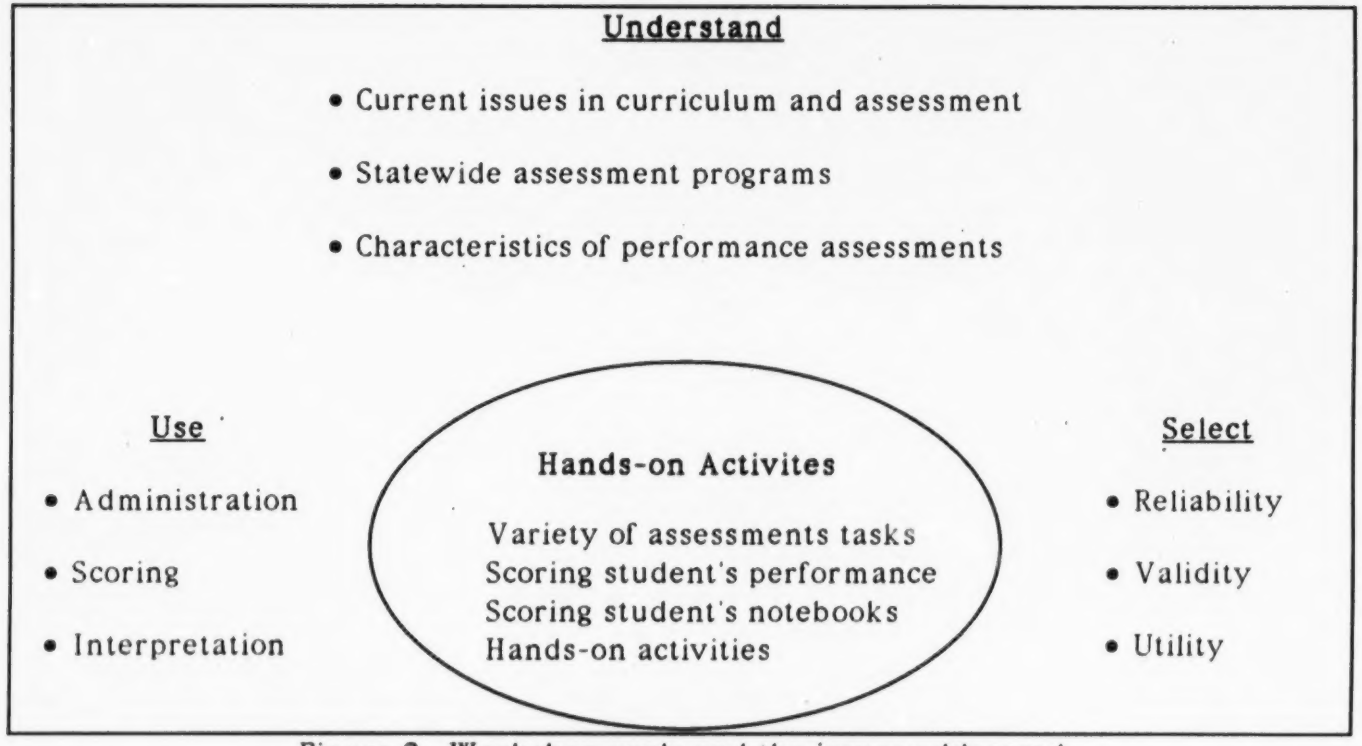

Figure 2. Workshop goals and the issues addressed.

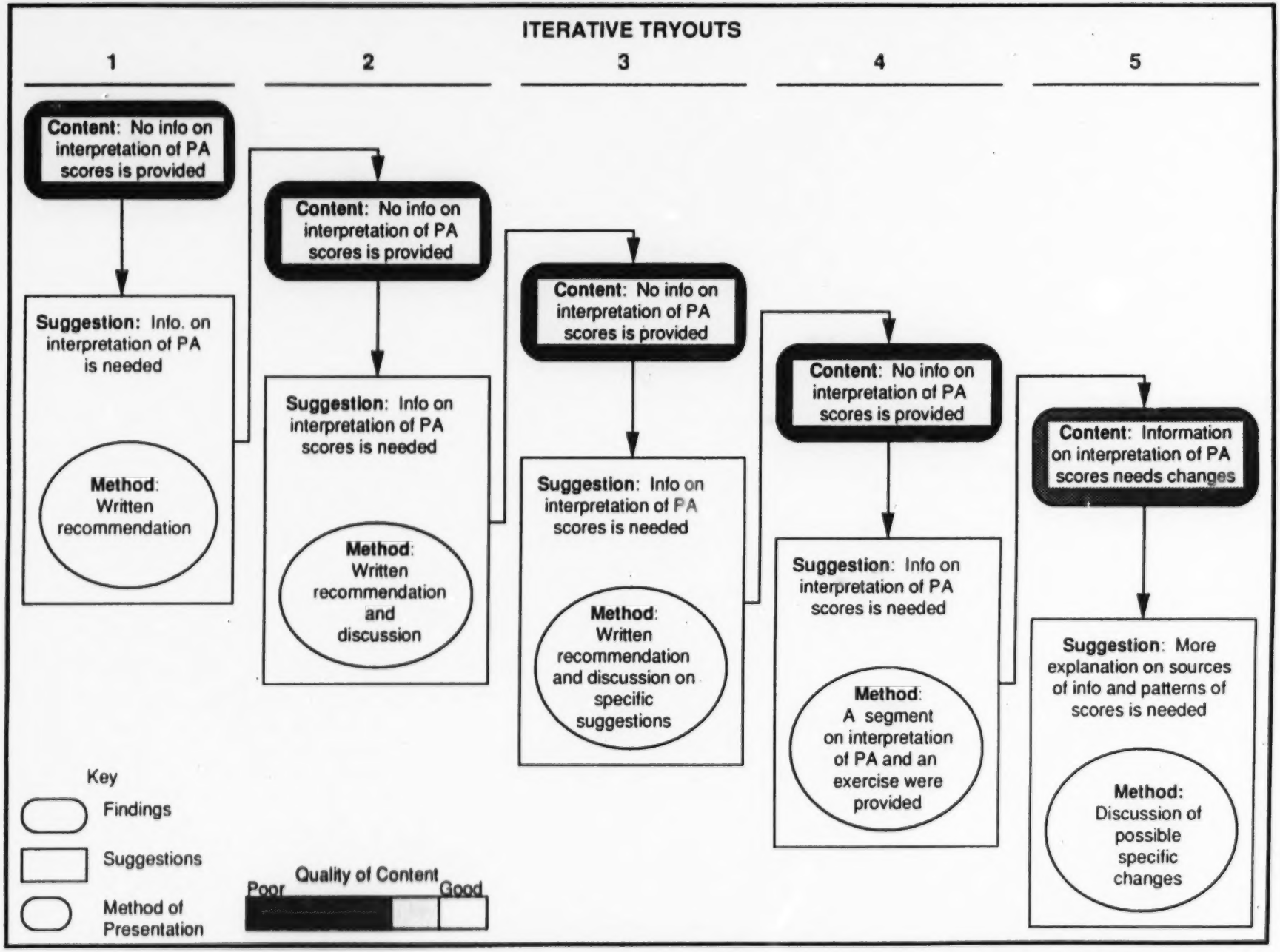
ITERATIVE TRYOUTS
1
2
3
4
5
Content: No info on interpretation of PA scores is provided
Suggestion: Info. on interpretation of PA is needed
Method: Written recommendation
Content: No info on interpretation of PA scores is provided
Suggestion: Info on interpretation of PA scores is needed
Method: Written recommendation and discussion
Content: No info on interpretation of PA scores is provided
Suggestion: Info on interpretation of PA scores is needed
Method: Written recommendation and discussion on specific suggestions
Content: No info on interpretation of PA scores is provided
Suggestion: Info on interpretation of PA scores is needed
Method: A segment on interpretation of PA and an exercise were provided
Content: Information on interpretation of PA scores needs changes
Suggestion: More explanation on sources of info and patterns of scores is needed
Method: Discussion of possible specific changes
Key
Findings
Suggestions
Method of Presentation
Quality of Content
Poor
Good

Figure 3. Development of the content on use of performance assessment–interpretation of scores, across tryouts.

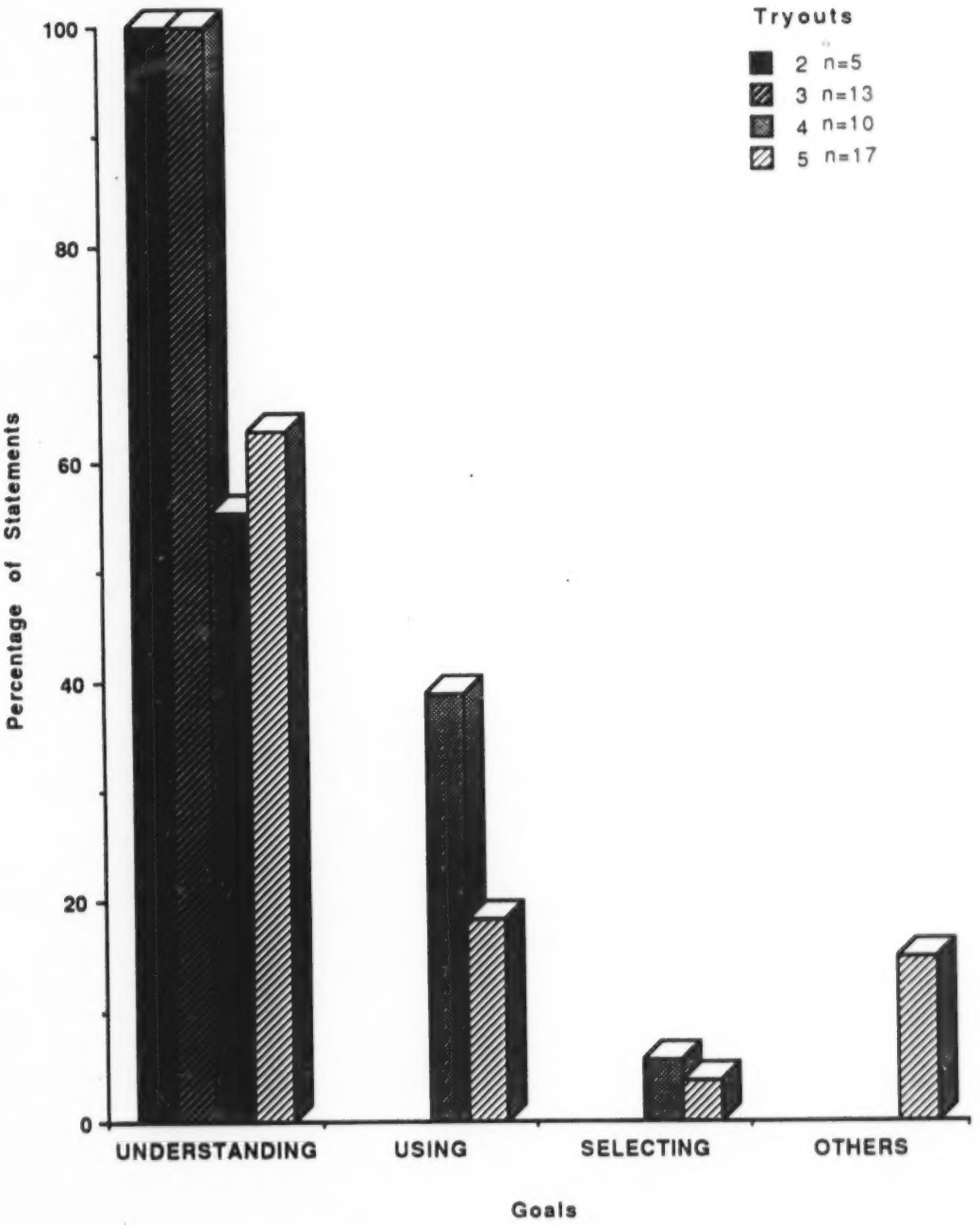

Figure 4. Participants' perceptions of workshop goals.

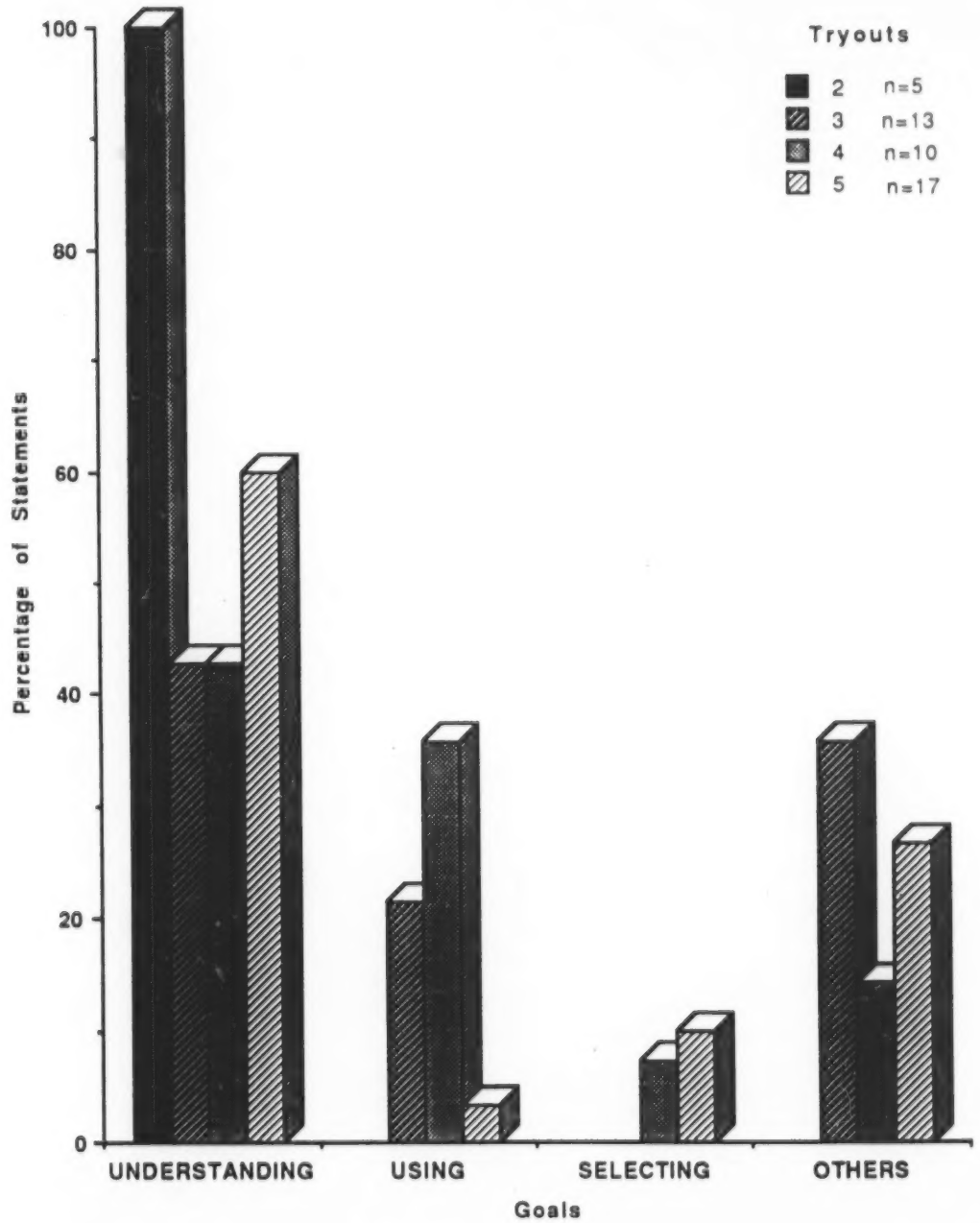

Figure 5. Participants' statements about benefits gained from the workshop by goal.

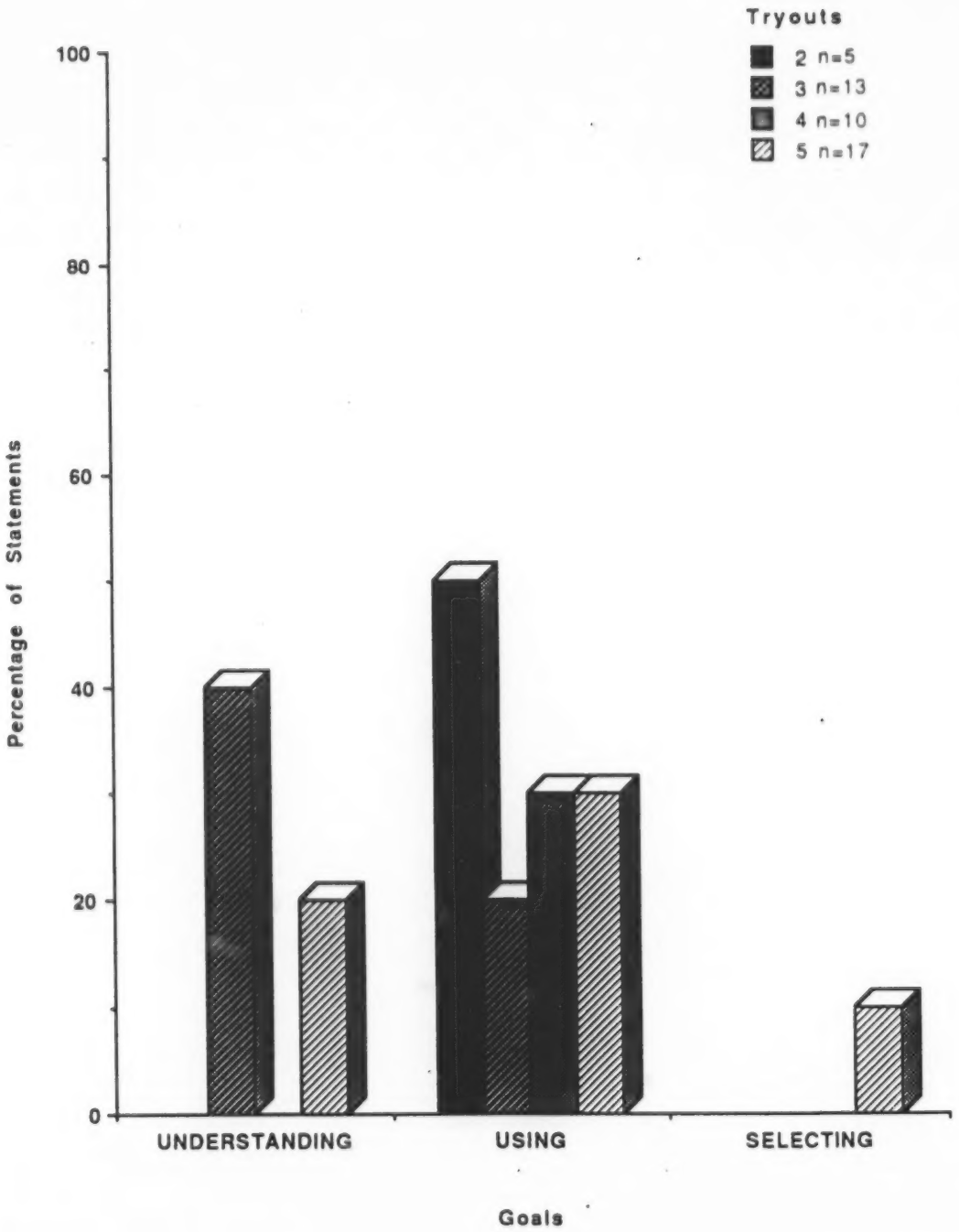

Figure 6. Participants' comments for improvement of workshop content by goal.

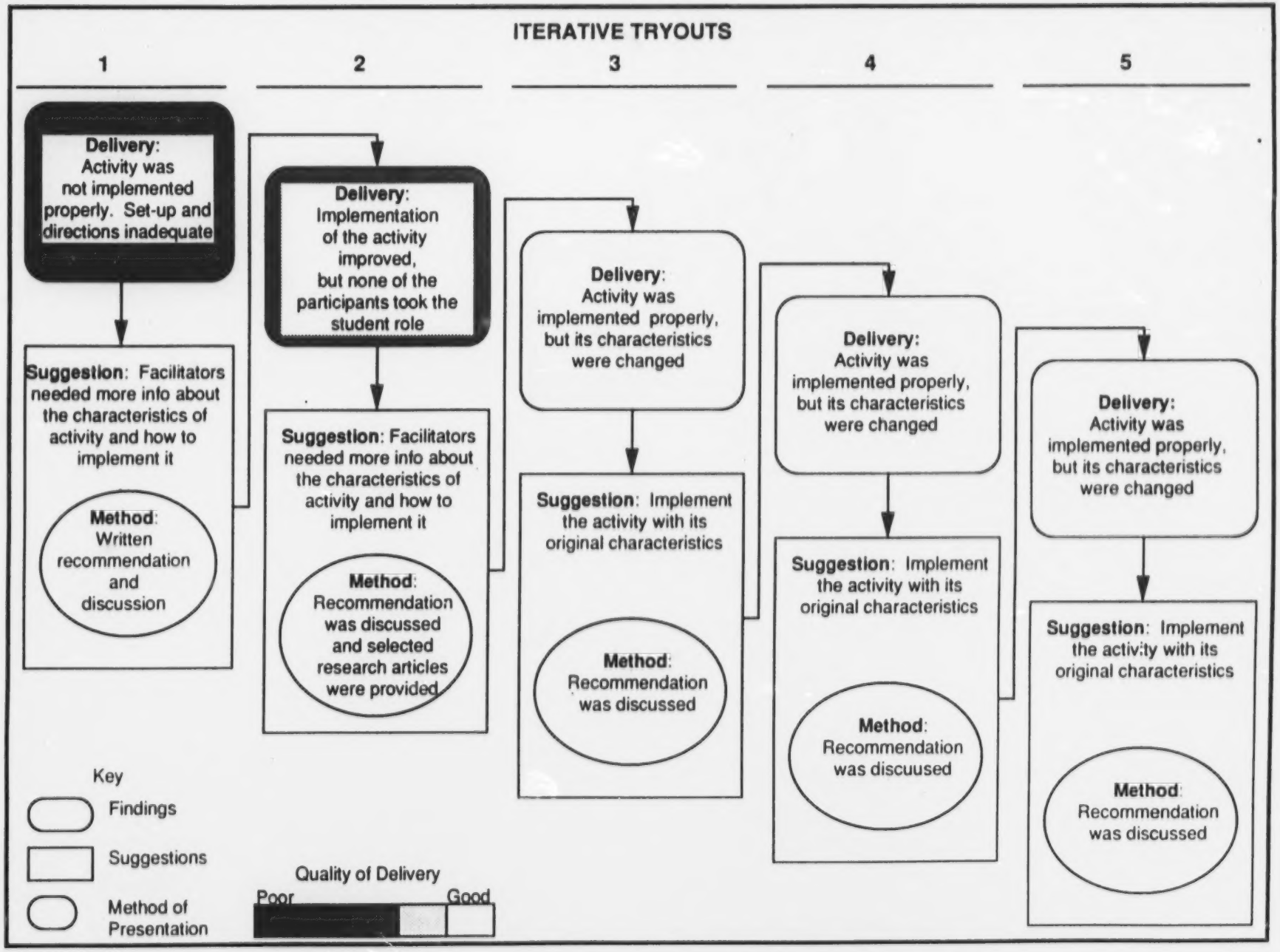
ITERATIVE TRYOUTS
1
2
3
4
5
Delivery: Activity was not implemented properly. Set-up and directions inadequate
Suggestion: Facilitators needed more info about the characteristics of activity and how to implement it
Method: Written recommendation and discussion
Delivery: Implementation of the activity improved, but none of the participants took the student role
Suggestion: Facilitators needed more info about the characteristics of activity and how to implement it
Method: Recommendation was discussed and selected research articles were provided
Delivery: Activity was implemented properly, but its characteristics were changed
Suggestion: Implement the activity with its original characteristics
Method: Recommendation was discussed
Delivery: Activity was implemented properly, but its characteristics were changed
Suggestion: Implement the activity with its original characteristics
Method: Recommendation was discuused
Delivery: Activity was implemented properly, but its characteristics were changed
Suggestion: Implement the activity with its original characteristics
Method: Recommendation was discussed
Key
Findings
Suggestions
Method of Presentation
Quality of Delivery
Poor
Good

Figure 7. Delivery of the Paper Towels activity across the tryouts

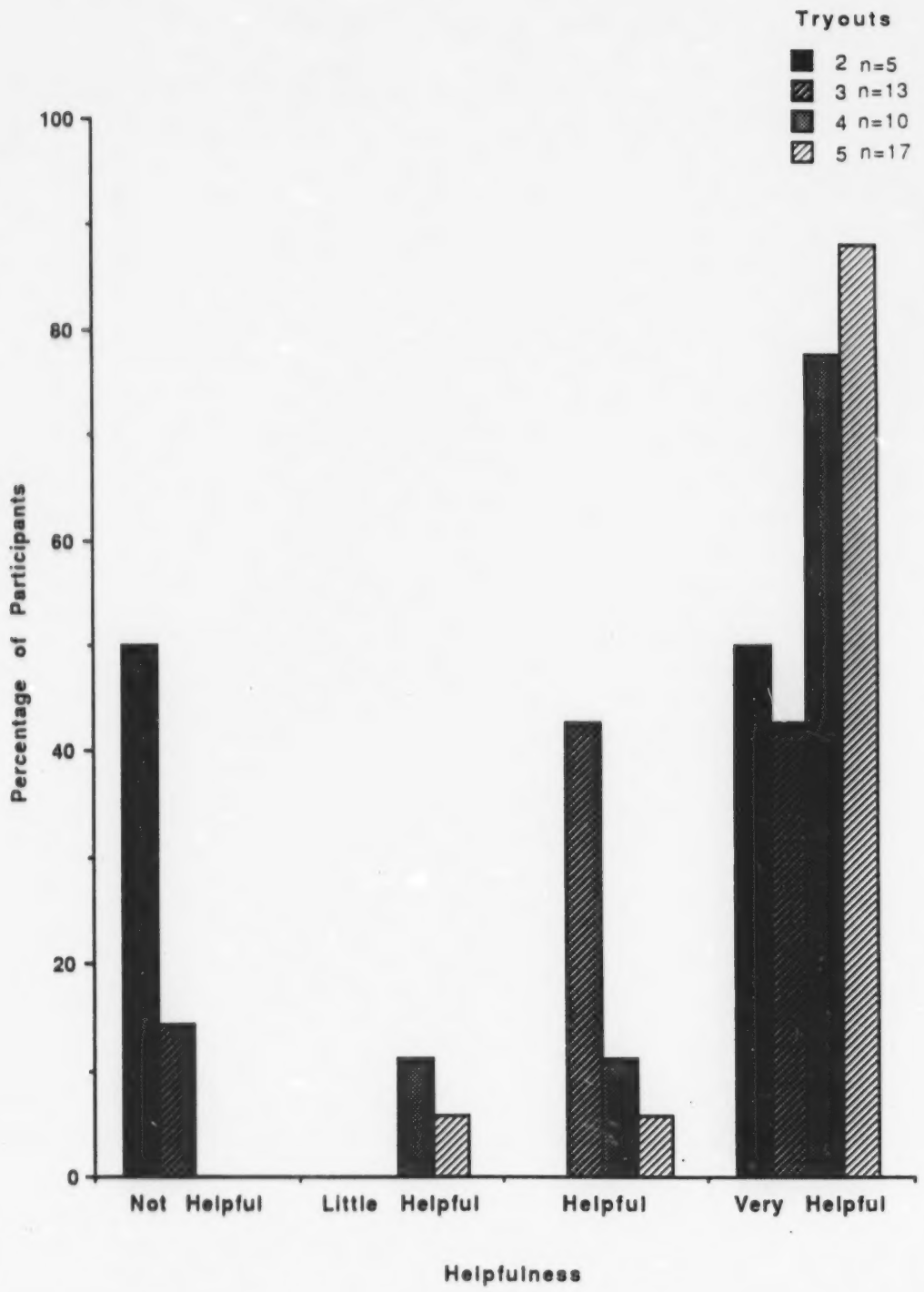

Figure 8. Participants' perceptions of helpfulness of the Paper Towels activity in understanding, using, selecting performance assessments.

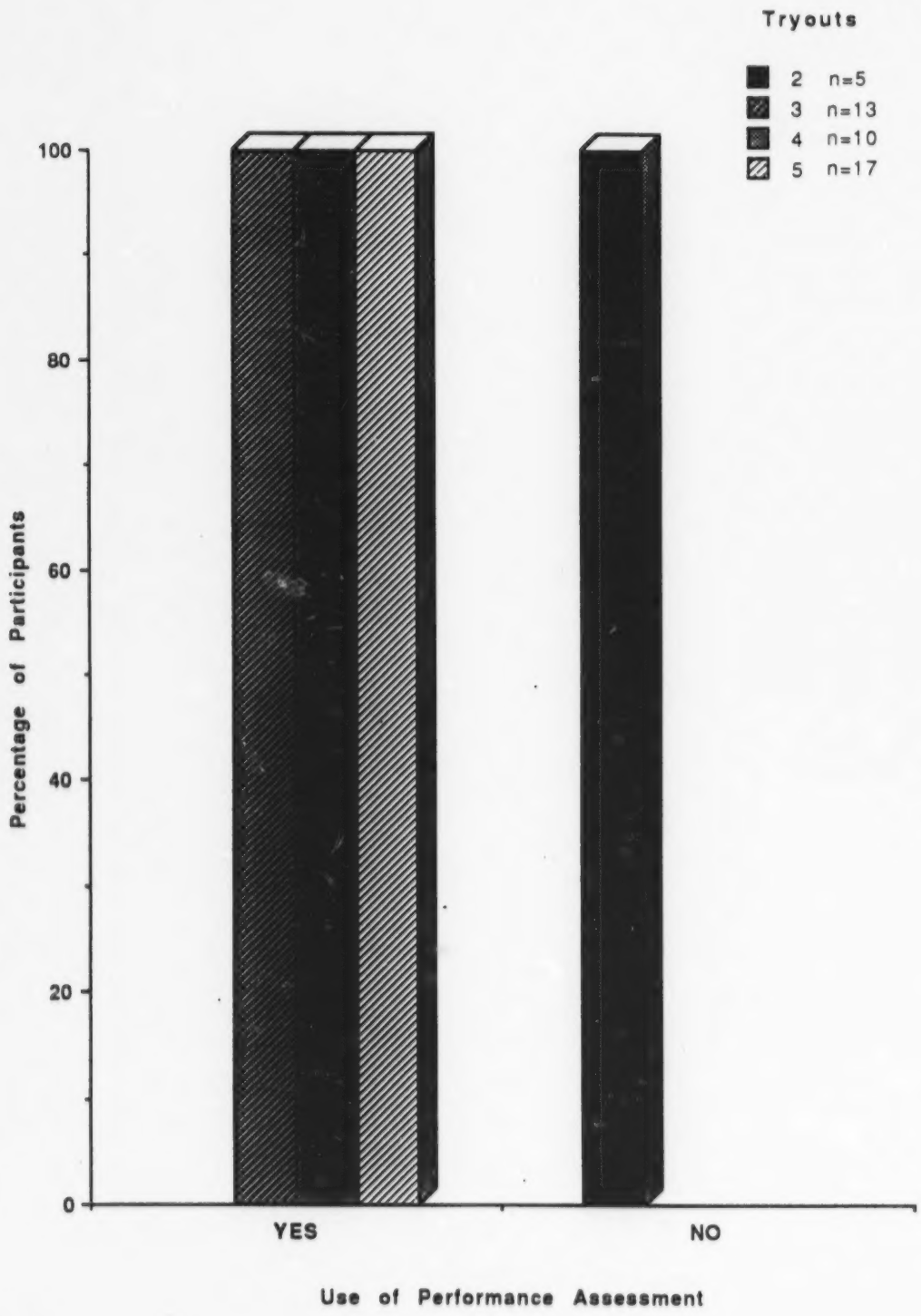

Figure 9. Participants' perceptions of their ability to use performance assessments in their class.

Appendix I

Content by Goal Across Tryouts

UNDERSTAND

TRYOUTS	1	2	3	4	5
Curriculum reform	• Current climate of assessments • Current changes in curriculum • Current reform and assessments	• Current curriculum reform • Current assessment practices • Needed reform in assessment	• Current assessment practices • Current curriculum reform • Needed reform in assessment	• Current assessment practices • Current curriculum reform • Needed reform in assessment	• Current assessment practices • Current curriculum reform • Needed reform in assessment
Statewide assessment programs		• California Assessment Program	• California Assessment Program	• California Assessment Program	• California Assessment Program
Characteristics of performance assessments	• Definition • Purposes of assessments • Varieties of performance assessments: 1) Paper Towels 2) Electric Mysteries • Variety of scoring forms: 1) Paper Towels scoring form 2) Electric Mysteries scoring form	• Definition • Purposes of assessments • Varieties of performance assessments: 1) Variety of assessments tasks 2) Paper Towels • Variety of scoring forms: 1) Paper Towels scoring form	• Definition • Purposes of assessments • Varieties of performance assessments: 1) Variety of assessments tasks 2) Paper Towels 3) Mystery Powders 4) Leaves • Variety of scoring forms: 1) Paper Towels scoring form 2) Mystery Powders scoring form 3) Rubric-Leaves scoring form	• Definition • Purposes of assessments • Varieties of performance assessments: 1) Variety of assessments tasks 2) Paper Towels 3) Mystery Powders 4) Leaves • Variety of scoring forms: 1) Paper Towels scoring form 2) Mystery Powders scoring form 3) Rubric-Leaves scoring form	• Definition • Purposes of assessments • Varieties of performance assessments: 1) Variety of assessments tasks 2) Paper Towels 3) Mystery Powders 4) Leaves • Variety of scoring forms: 1) Paper Towels scoring form 2) Mystery Powders scoring form 3) Rubric-Leaves scoring form

USE

TRYOUTS	1	2	3	4	5
Administration	• Embedded assessments		• Embedded assessments • Discussion of difficulties in administration of PA	• Embedded assessments • Discussion of difficulties in administration of PA	• Embedded assessments • Discussion of difficulties in administration of PA
Scoring	• Scoring students' performance in Paper Towels assessment • Scoring students' notebooks of Paper Towels assessment	• Scoring students' performance in Paper Towels assessment	• Scoring students' performance in Paper Towels assessment • Scoring students' notebooks of Paper Towels assessment • Scoring students' notebooks in Mystery Powders assessment • Scoring students' notebooks of Leaves assessment	• Scoring students' performance in Paper Towels assessment • Scoring students' notebooks of Paper Towels assessment • Scoring students' notebooks in Mystery Powders assessment • Scoring students' notebooks of Leaves assessment	• Scoring students' performance in Paper Towels assessment • Scoring students' notebooks of Paper Towels assessment • Scoring students' notebooks in Mystery Powders assessment • Scoring students' notebooks of Leaves assessment
Interpretation					• Strategy to interpret performance assessment scores

SELECTION

TRYOUTS	1	2	3	4	5
Reliability	• Interrater reliability	• Interrater reliability	• Interrater reliability	• Interrater reliability	• Interrater reliability
Validity		• Criterion validity • Content validity	• Criterion validity: Three examples of this type of validity • Content validity	• Criterion validity: Three examples of this type of validity • Content validity	• Criterion validity: Three examples of this type of validity • Content validity
Utility			• Embedded assessments • Discussion on usefulness of assessment tasks and information it produces	• Embedded assessments • Discussion on usefulness of assessment tasks and information it produces • Exercise on selection of performance assessment	• Discussion on usefulness of assessment tasks and information it produces • Exercise on selection of performance assessment